Peut-on utiliser les étiqueteurs morphosyntaxiques pour améliorer la transcription automatique ?

Stéphane Huet, Guillaume Gravier, Pascale Sébillot

Peut-on utiliser les étiqueteurs morphosyntaxiques pour améliorer la transcription automatique ?

Stéphane Huet, Guillaume Gravier, Pascale Sébillot

IRISA
Campus de Beaulieu, F-35042 Rennes Cedex, France
shuet@irisa.fr, ggravier@irisa.fr, sebillot@irisa.fr

ABSTRACT

The aim of the paper is to study the interest of part-of-speech (POS) tagging to improve speech recognition. We first evaluate the part of misrecognized words that can be corrected using POS information ; an analysis of a short extract proves that an absolute decrease of the word error rate by 1.1 % can be expected. We also demonstrate quantitatively that traditional POS taggers are reliable when applied to spoken corpus, including automatic transcriptions. This new result enables us to effectively use POS tag knowledge to improve, in a postprocessing stage, the quality of transcriptions, especially correcting agreement errors.

1. INTRODUCTION

Les systèmes de transcription utilisent globalement peu de connaissance sur le langage pour décoder la parole, se limitant en cela bien souvent au seul apprentissage des probabilités de successions de mots sur un corpus. Au vu du matériau manipulé, de la langue naturelle, il semble pourtant que des informations linguistiques supplémentaires devraient permettre d'améliorer la qualité de la transcription. Certains modèles de langage (ML) ont d'ailleurs déjà été conçus en intégrant des connaissances sur la structure syntaxique des groupes de souffle [2], sur les thèmes abordés par le document à transcrire [6] ou encore sur les parties du discours (appelées aussi *POS* pour *part of speech*) [7]. Une POS correspond à une propriété grammaticale d'un mot ou groupe de mots dans une phrase donnée (*e.g.* noms, verbes, prépositions, conjonctions, *etc.*), souvent accompagnée d'informations morphologiques (genre, nombre, conjugaison, *etc.*). La connaissance de ces catégories est généralement prise en compte au sein des ML à l'aide de modèles N-classes [1]. Si $\mathcal{C}_i$ représente l'ensemble des POS c_i auxquelles peut appartenir un mot w_i, le calcul des probabilités de la séquence de mots $w_1^n = w_1, ..., w_n$ s'effectue selon

$$P(w_1^n) \approx \sum_{c_1 \in \mathcal{C}_1, ..., c_n \in \mathcal{C}_n} \prod_{i=1}^{n} P(w_i|c_i) P(c_i|c_{i-N+1}^{i-1}) \quad (1)$$

L'interpolation des modèles N-classes avec des modèles N-grammes conduit généralement à une baisse négligeable du taux d'erreur sur les mots de la transcription. Diverses améliorations ont donc été envisagées. Il a ainsi été proposé d'estimer la probabilité en considérant que les POS associées aux mots w_i à reconnaître font partie intégrante de la sortie de la transcription et ne sont plus un simple résultat intermédiaire [5]. Cette approche évalue les probabilités à l'aide d'un mode de calcul plus précis que celui de (1) mais conduit à une augmentation importante du nombre d'événements à considérer.

Dans ces diverses approches, les POS servent à construire des ML intervenant au cours du processus de transcription. Or ce type d'utilisation apporte un gain limité par rapport aux ML N-grammes de mots. Nous proposons donc, dans cet article, d'étudier la possibilité d'utiliser les étiquettes POS en aval de la transcription, pour sélectionner la meilleure hypothèse parmi plusieurs proposées par le système de reconnaissance. La première étape de notre travail a consisté à déterminer la proportion des erreurs de transcription corrigeable par la connaissance des POS. Celle-ci étant importante, nous avons cherché à évaluer la capacité des méthodes automatiques à étiqueter des transcriptions. Les étiqueteurs sont en effet conçus à l'origine pour des documents écrits, dont certaines caractéristiques sont très différentes d'un corpus oral, d'autant plus si celui-ci comporte des erreurs de transcription. Nos différentes évaluations ayant montré l'aptitude des étiqueteurs, nous avons mené de premières expérimentations pour tester l'utilisation des POS en post-traitement du système de transcription. Le plan de la suite de l'article suit les différentes étapes de notre démarche.

2. TYPOLOGIE DES ERREURS DE TRANSCRIPTION

Afin d'évaluer l'apport potentiel des POS pour la transcription, nous avons étudié en détail un court extrait de transcription automatique, en cherchant à connaître la part des erreurs corrigeables par cette seule connaissance.

Le système de reconnaissance utilisé dans nos expérimentations, développé par l'IRISA et l'ENST pour la campagne ESTER, permet de produire un graphe de mots en trois passes[1]. Un premier graphe de mots est généré avec des modèles acoustiques hors-contexte et un ML trigramme. La deuxième passe utilise des modèles contextuels pour réévaluer les 1 000 meilleurs chemins extraits du graphe de la première passe à l'aide d'un ML 4-gramme. Enfin, la troisième passe, similaire à la deuxième après adaptation au locuteur des modèles contextuels, permet de générer un graphe d'hypothèses. Afin de s'affranchir des problèmes de segmentation, nous considérons dans ce travail une segmentation manuelle en groupes de souffle basée sur la détection de pauses silencieuses. Nous examinons ici un extrait de 6 500 mots d'une émission d'information sur France-Inter, issue du corpus ES-

[1]Nous tenons à remercier François Yvon pour nous avoir fourni le ML et le lexique étiqueté.

TER [4]. Le taux d'erreur sur cet extrait est de 17,8%.

Parmi les erreurs de reconnaissance que nous y avons constatées, trois groupes se détachent. Certaines erreurs correspondent à un « dérapage » du système, généralement dû soit à une mauvaise acoustique, soit à une mauvaise reconnaissance d'entités nommées. Ces erreurs semblent hors d'atteinte de la correction susceptible d'être apportée par les POS. Heureusement, elles ne concernent qu'une part très restreinte de l'extrait analysé. Le deuxième ensemble correspond à des groupes de souffle agrammaticaux (Fig. 1). L'agrammaticalité est notamment causée par les mots grammaticaux « *a* », « *à* », « *de* », « *que* » ou encore « *et* », qui sont parfois absents, ou présents de manière inopinée dans les hypothèses de transcription. On retrouve également des fautes de temps et de mode des verbes, le présent et l'indicatif étant souvent privilégiés. Parmi ces erreurs, certaines semblent corrigeables puisque l'étiquetage des groupes de souffle peut conduire à des séquences de POS aberrantes, comme l'apparition de trois prépositions consécutives. Ce critère est néanmoins à prendre avec précaution, à cause des répétitions présentes dans la langue parlée. Le troisième groupe est formé d'erreurs très vraisemblablement corrigeables grâce aux POS, à savoir les fautes d'accord en genre et en nombre et les confusions entre infinitif et participe passé. Ces erreurs sont particulièrement nombreuses puisqu'elles concernent un groupe de souffle sur sept. Parmi elles, 70 sont rectifiables sans avoir à examiner de dépendances entre des groupes de souffle consécutifs (Fig. 1), et les corriger représenterait une baisse absolue de 1,1% du taux d'erreur. Au travers de l'exposé des principales erreurs de décodage, il apparaît donc que les POS constituent une source d'information intéressante pour améliorer la qualité de la transcription.

3. COMPORTEMENT DES ÉTIQUETEURS

La section précédente a montré l'intérêt des étiquettes POS pour corriger des erreurs de transcription en se focalisant sur des successions de POS possibles. Pour pouvoir utiliser cette technique, il faut cependant que les étiqueteurs fonctionnent de manière fiable sur des corpus oraux produits par des annotateurs ou obtenus par des systèmes de reconnaissance. C'est cette propriété que nous cherchons à évaluer ici.

Le rôle des étiqueteurs morphosyntaxiques est d'associer à chaque mot ou groupe de mots de la séquence à étudier l'étiquette catégorielle la plus probable. Ces outils sont ordinairement appliqués sur des corpus écrits. Pour les évaluer quantitativement, un texte est étiqueté manuellement par des annotateurs et les étiquettes sont comparées une à une avec celles proposées par la méthode automatique. Comparativement aux corpus écrits, les corpus oraux, transcrits par des annotateurs, ont été peu étudiés [8]. La production orale présente des caractéristiques, telles que les reprises ou les répétitions, qui sont susceptibles de compliquer l'opération d'étiquetage. L'étiquetage de la transcription automatique de la parole planifiée est une tâche rendue plus complexe encore par le fait que le texte est segmenté en groupes de souffle et non en phrases, et ne contient ni ponctuations, ni majuscules (cas du vocabulaire de notre système de transcription).

De manière à faciliter l'utilisation des POS pour décoder la parole, nous avons fait le choix de construire notre propre étiqueteur morphosyntaxique. La suite de cette section décrit le protocole utilisé pour construire cet étiqueteur, avant d'évaluer son comportement sur de la parole transcrite. Nous avons examiné la qualité de l'étiquetage produit sur un corpus de test et l'avons comparée avec les résultats obtenus avec un étiqueteur qui fait référence pour le français.

3.1. Constitution d'un étiqueteur

Les étiqueteurs conçus pour l'écrit utilisent des règles linguistiques ou extraient automatiquement l'information statistique contenue dans de grands volumes de données. Dans la mesure où les programmes basés sur des calculs statistiques conduisent à des résultats satisfaisants pour l'écrit et ne nécessitent pas l'écriture manuelle de nombreuses règles contextuelles, nous avons construit notre étiqueteur en utilisant exclusivement des méthodes statistiques.

Pour ce faire, nous avons constitué un corpus d'apprentissage de 200 000 mots représentant un extrait d'une durée de 16 heures du corpus ESTER. Les transcriptions manuelles, contenant à l'origine des majuscules et des ponctuations, ont été étiquetées par le logiciel Cordial[2]. Le résultat a été vérifié manuellement, puis nous avons ôté toutes les majuscules et les marques de ponctuation pour qu'il soit cohérent avec la forme du texte produit par notre système de reconnaissance. Nous avons utilisé un lexique de prononciations étiqueté afin de connaître les POS possibles pour chaque mot. Le choix des étiquettes morphosyntaxiques a été fait de manière à discriminer le genre et le nombre des adjectifs et des noms, et le temps et le mode des verbes, ce qui conduit à un jeu de 80 étiquettes différentes. Cet ensemble d'étiquettes est très proche de celui proposé dans les grammaires scolaires et est directement inspiré de celui utilisé par Cordial.

Notre étiqueteur morphosyntaxique se base sur un modèle N-classe pour trouver la séquence d'étiquettes qui maximise le produit dans (1) pour une séquence de mot w_1^n. Des réglages sur un corpus de développement nous ont conduit à choisir un ordre $N = 7$ et un lissage de Kneser-Ney non modifié. De façon à évaluer l'impact de la segmentation sur la qualité de l'étiquetage, nous avons procédé à deux apprentissages différents, en segmentant le corpus d'apprentissage par phrases puis par groupes de souffle.

3.2. Évaluation de l'étiquetage

Afin d'avoir une mesure quantitative de la qualité de l'étiquetage sur des transcriptions produites manuellement (REF), segmentées en phrases ou en groupes de souffle, ou produites automatiquement (HYP) par le système de reconnaissance, nous avons étiqueté manuellement une émission d'information de France-Inter d'une heure, constituée de 11 300 mots, que nous désignerons par GOLD. L'étiquetage automatique de REF a été évalué en dénombrant le nombre d'étiquettes en commun avec GOLD. La mesure de la qualité de l'étiquetage a été plus problématique pour HYP, pour lequel nous avons mesuré un taux d'erreur de transcription de 22%, puisque les mots ne sont pas identiques avec ceux du GOLD. Il a ainsi été impossible de constituer un étiquetage de référence pour HYP dans la mesure où il n'existe pas de POS cohérentes pour les mots des groupes de souffle agrammaticaux. Nous

[2] Version 8.1 distribuée par la société *Synapse Développement*.

Hypothèse agrammaticale
REF: bush ** SAIT donc QU' il faudra coopérer
HYP: bush s' EST donc ** il faudra coopérer
Erreur d'accord
REF: c' est un monstre injuste envers sa soeur si DÉVOUÉE
HYP: c' est un monstre injuste envers sa soeur si DÉVOUÉ

FIG. 1: Exemples d'erreurs dans les groupes de souffle

donnons donc deux mesures de la qualité de l'étiquetage de HYP : le pourcentage de mots correctement reconnus et étiquetés parmi le nombre total de mots du GOLD, et le pourcentage de mots correctement reconnus et étiquetés parmi le nombre de mots bien reconnus dans HYP (donné entre parenthèses dans le tableau 1).

Les résultats obtenus par notre étiqueteur sur les corpus de test sont présentés dans les deux premières lignes du tableau 1, en effectuant toutes les compositions possibles en ce qui concerne la segmentation du corpus d'apprentissage et des corpus de test. Ils établissent que l'étiquetage produit est bon dans l'ensemble, y compris pour les transcriptions automatiques dont les erreurs de reconnaissance peuvent venir perturber l'étiquetage des mots correctement transcrits. Ces résultats sont relativement surprenants dans la mesure où nous n'avons pas introduit de méthodes spécifiques pour traiter les particularités de la langue orale, si ce n'est d'utiliser un corpus oral pour paramétrer l'étiqueteur. Cette robustesse des étiqueteurs sur la langue parlée s'explique cependant par le fait que les étiquettes sont attribuées en exploitant des informations de manière locale. Il apparaît en outre que l'apprentissage à partir d'une segmentation en groupes de souffle fournit les meilleurs résultats, ce qui nous a conduit à privilégier ce mode de segmentation par la suite.

De plus, en examinant les fautes commises dans l'attribution des POS, nous avons constaté que certaines pouvaient être considérées comme acceptables. Ainsi, les distinctions entre les POS « participe passé » et « adjectif » sont dans la grande majorité des cas discutables. Nous avons également constaté de nombreuses erreurs dues à la mauvaise tokenisation de notre étiqueteur. Ainsi, alors que le GOLD avait étiqueté respectivement « *états-unis* » et « *alors que* » comme nom propre et conjonction de subordination, l'étiquetage automatique a conduit à reconnaître d'une part « *états* » comme nom commun, « *unis* » comme adjectif et, d'autre part, « *alors* » comme adverbe et « *que* » comme conjonction de subordination. Sur les 966 erreurs observées lors de l'étiquetage de REF segmenté par groupes de souffle, 42 sont dues à des confusions entre participe passé et adjectif, 216 à des erreurs de tokenisation, 124 à des confusions entre nom commun et nom propre et 10 à des mots inconnus par l'étiqueteur.

Nous avons en outre comparé les performances de notre étiqueteur à celles de Cordial, vraisemblablement meilleur étiqueteur disponible pour le français écrit, qui a déjà donné de bons résultats sur un corpus de parole [8]. La dernière ligne du tableau 1 présente ses résultats sur le corpus de test. L'examen de ce tableau établit que notre étiqueteur a des résultats comparables, voire meilleurs que Cordial. On peut d'ailleurs constater que Cordial se comporte moins bien qu'habituellement, ses scores sur de l'écrit étant généralement supérieurs à 95%. Ceci s'explique par la nature particulière de la transcription automatique, pour laquelle il n'a pas été spécifiquement conçu. L'absence de majuscules est particulièrement problématique dans la mesure où le logiciel s'appuie sur cet indice pour détecter les noms propres. En ignorant toutes les erreurs dues à une confusion entre nom commun et nom propre, le pourcentage d'étiquettes bien attribuées monte à 93,52% pour le corpus de test segmenté par groupes de souffle, alors que, suivant le même critère et sur les mêmes données de test, les performances de notre étiqueteur ne progressent qu'à 92,55%.

Cette série d'expérimentations montre que l'étiquetage des transcriptions automatiques est fiable, ce qui n'était encore qu'une hypothèse auparavant. Notre étiqueteur conduit à des résultats qui nous permettent de l'envisager pour calculer un score sur la qualité du décodage. La section suivante présente des résultats préliminaires sur l'utilisation de la connaissance des POS en post-traitement de la transcription.

4. APPORT DE L'ÉTIQUETAGE À LA TRANSCRIPTION

De manière à exploiter la connaissance des POS au cours du décodage de la parole, nous avons employé notre étiqueteur pour attribuer un score à chaque hypothèse donnée sur un groupe de souffle. Chaque hypothèse candidate w_1^n est étiquetée par la séquence de POS c_1^n la plus probable, avant d'être évaluée par la fonction de score suivante :

$$lp = \log P(c_1^n) = \sum_{i=1}^{n} \log P(c_i | c_{i-N+1}^{i-1}) \qquad (2)$$

Ce score vise à réévaluer la liste des meilleures hypothèses produites par le système de reconnaissance pour chaque groupe de souffle.

Afin de valider notre approche, nous avons testé dans un premier temps le comportement sur les 70 erreurs d'accord que nous avions repérées (*cf.* section 2). Pour chacun des groupes de souffle contenant l'une de ces erreurs, nous avons établi le score pour trois versions : la transcription de référence (REF), la transcription automatique (HYP) et la transcription automatique où seules les erreurs d'accord sont corrigées (COR) (Fig. 2). Nous espérons ainsi que la succession d'étiquettes obtenues sur REF et COR sera plus probable que sur HYP.

Nous avons constaté sur les 63 groupes de souffle analysés que le score était meilleur sur COR que sur HYP pour 46 d'entre eux et meilleur sur REF que sur HYP pour 41 d'entre eux. Ces résultats établissent ainsi que, dans une majorité des cas, lp conduit à une correction des fautes d'accord et est donc susceptible d'apporter un gain au niveau du taux d'erreur de reconnaissance.

TAB. 1: Évaluation des étiqueteurs (en pourcentages)

	REF / phrase	REF / groupe de souffle	HYP
corpus d'app / phrase	91,42	91,09	72,60 (91,83)
corpus d'app / groupe de souffle	91,50	91,42	72,99 (92,32)
Cordial	88.69	88,61	70,75 (89,48)

```
REF: à  L' AMÉNAGER avant qu' elle ne soit DÉTRUITE
COR: à  LA MÉNAGER  avant qu' elle ne soit DÉTRUITE
HYP: à  LA MÉNAGER  avant qu' elle ne soit DÉTRUIT
```

FIG. 2: Versions à évaluer pour un même groupe de souffle

Nous avons utilisé ce score pour réordonner la liste des 100 meilleures hypothèses produites pour 4 heures d'émissions d'informations en français. Le taux d'erreur sur les mots donné par un oracle sur cette liste est de 14,2 % pour un taux initial de 21,6 %. En réordonnant la liste en utilisant lp, le taux augmente de manière significative de 21,6 % à 26,2 %. Nous avons donc décidé de combiner le score sur les POS avec le score acoustique et le score du ML.

La reconnaissance de la parole est en pratique habituellement exprimée comme une recherche de w_1^n à partir de l'entrée acoustique y_1^n. Pour introduire lp, nous modifions le critère de sélection de w_1^n par

$$\hat{w}_1^n = \arg\max_{w_i^n} [\log P(y_1^n|w_1^n) + \alpha \log P(w_1^n) + \beta \log P(c_1^n) + \gamma n] \quad (3)$$

où α est le facteur d'échelle du ML et γ est la pénalité d'insertion d'un mot. $P(w_1^n)$ est calculée par un ML 4-gramme sur les mots, tandis que $P(c_1^n)$ est déterminée par un ML 7-gramme sur les POS.

Nous avons observé une légère diminution du taux d'erreur à 21,4 % avec cette méthode. Nous avons également remarqué que généralement les erreurs d'accord ont été corrigées. Par exemple, le groupe de souffle initialement transcrit par « *le messin disputent aujourd'hui* » a été correctement rectifié par « *le messin dispute aujourd'hui* ». Toutefois, quelques erreurs apparaissent comme la transcription, correcte au départ, « *les visages de Jacques Chirac et Jean-Marie Le Pen apparaissent* » qui a été modifiée en « *les visages de Jacques Chirac et Jean-Marie Le Pen apparaît* ».

La connaissance des POS apporte donc une information réduite par rapport au ML basé sur les mots, bien que les deux méthodes soient complémentaires comme le montre la réduction du taux d'erreur. Notre approche semble en outre un peu plus performante que celle des ML N-classes classiques puisqu'en effectuant une combinaison linéaire de ce type de modèle avec un ML 4-gramme, et ce, en utilisant le même jeu d'étiquettes que précédemment, le taux d'erreur n'a pu être réduit en dessous de 21,6 %.

5. PERSPECTIVES

Dans cet article, nous avons d'une part montré l'intérêt de la connaissance des étiquettes POS pour corriger des erreurs de transcription de parole pour le français et avons, d'autre part, prouvé quantitativement que les étiqueteurs pouvaient réellement être utilisés sur des corpus oraux transcrits manuellement ou automatiquement, ce qui rend effectivement possible l'exploitation des POS pour améliorer les transcriptions. De premières expériences ont montré que les étiquettes POS pouvaient corriger des fautes d'accord, même si cela se manifeste globalement par une diminution modeste du taux d'erreur sur les mots. Pour améliorer ces premiers résultats, au lieu d'opérer sur les N meilleures hypothèses du système de transcription, nous prévoyons de réévaluer tous les homophones de la meilleure hypothèse trouvée [3]. En outre, nous souhaitons étudier l'influence d'autres jeux d'étiquettes POS sur la qualité de la transcription.

RÉFÉRENCES

[1] P.F. Brown, V.J. Della Pietra, P.V. deSouza, J.C. Lai, and R.L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–480, 1992.

[2] C. Chelba and F. Jelinek. Structured language modeling. *Computer Speech and Language*, 14(4):283–332, 2000.

[3] J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, and H. Schwenk. Where are we in transcribing French broadcast news? In *Proc. of Eurospeech*, 2005.

[4] G. Gravier, J.-F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait, and K. Choukri. ESTER, une campagne d'évaluation des systèmes d'indexation d'émissions radiophoniques. In *Actes des JEP*, 2004.

[5] P.A. Heeman. POS tags and decision trees for language modeling. In *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[6] S. Khudanpur and J. Wu. A maximum entropy language model to integrate n-grams and topic dependencies for conversational speech recognition. In *Proc. of ICASSP*, 1999.

[7] G. Maltese and F. Mancini. An automatic technique to include grammatical and morphological information in a trigram-based statistical language model. In *Proc. of ICASSP*, volume 1, 1992.

[8] A. Valli and J. Véronis. Étiquetage grammatical de corpus oraux : problèmes et perspectives. *Revue française de linguistique appliquée*, 4(2):113–133, 1999.